Engineered for Redundancy: How Engineered Systems Handle Hardware Failures

Presented by: Andy Colvin
Oracle Open World 2012
October 2, 2012

enkitec

About Me

- Working around Oracle since 1999
- Background in systems, network, database
- 6 years at Enkitec
- Working on Exadata for 2+ years
- I'm a little "door key"

enkitec

About Enkitec

- Oracle-Centric Consulting Firm
  - US
  - UK
- Extensive Exadata Practice
  - Installation
  - Migration
  - Performance Review
  - On Call Support
  - Education
- Booth 421 – Moscone South

enkitec

Talking About Redundancy

enkitec

Oracle's Engineered Systems

Exadata

Big Data Appliance

SPARC SuperCluster

Exalogic

Oracle Database Appliance

Exalytics

enkitec

Why Engineered Systems?

- One provider for hardware/software
- Fast deployment
- Redundant hardware/software

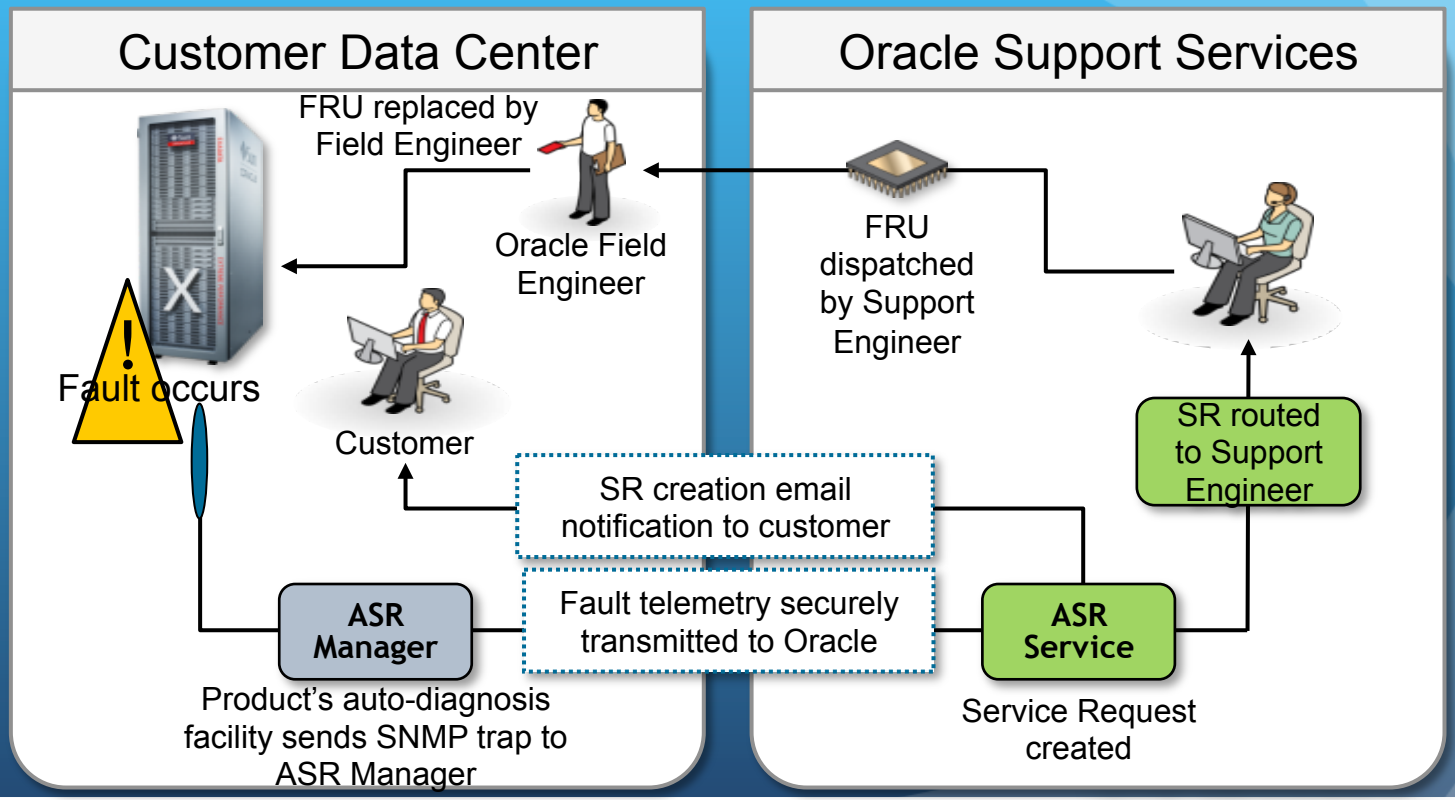
enkitec

Automatic Service Request

- Creates new SR when hardware fails
- Integrates with Oracle Enterprise Manager Ops Center
- Requires separate server running OEL/Solaris
- One-way communication to Oracle support

enkitec

Automatic Service Request



Automatic Service Request

ORNASRINTFC_WW@ORACLE.COM - April 8, 2012 11:49:46 AM GMT-05:00 [Customer Problem Description]

Problem Description: =ASR Alarm=
Automatic Service Request (ASR) Alarm
Generated: 2012-04-08 10:49:21
Severity : 2
Device : [redacted]
Eventcode: HALRT-02001

Event num: HALRT-02001

ASR: System hard disk failure

Hostname: dm03cel06-ilom.[redacted]
Product Type: SUN FIRE X4270 M2 SERVER
Summary:ASR: System hard disk failure
Description:Please review MOS note 1112994.1 by selecting "Related Articles"

sunHwTrapChassisId = [redacted]
sunHwTrapProductName = SUN FIRE X4270 M2 SERVER
sunHwTrapSuspectComponentName = SEAGATE ST32000SSSUN2.0T; Slot: 1
sunHwTrapFaultClass = NULL
sunHwTrapFaultCertainty = 0
sunHwTrapFaultMessageID = HALRT-02001
sunHwTrapFaultUUID = 5f93b935-a04f-4a63-84fe-faa25f4dcb25
sunHwTrapAssocObjectId = .0.0
sunHwTrapAdditionalInfo = Exadata Storage Server: dm03cel06.[redacted] Disk Serial Number: L78518

Please review MOS note 1112994.1 by selecting "Related Articles"
Extra information:-
Alerts received for this system in last 2 months (limit 10):
None

enkitec

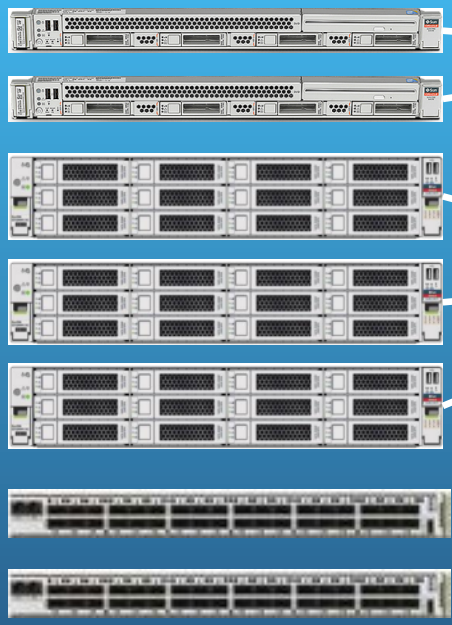
What Next?

What to do after the SR is created?

- Collect hardware diagnostics
  - ILOM snapshot – MOS Note #1448069.1
  - sundiag.sh – MOS Note #761868.1
- Replace with part from spares kit

enkitec

Redundant Hardware - Exadata



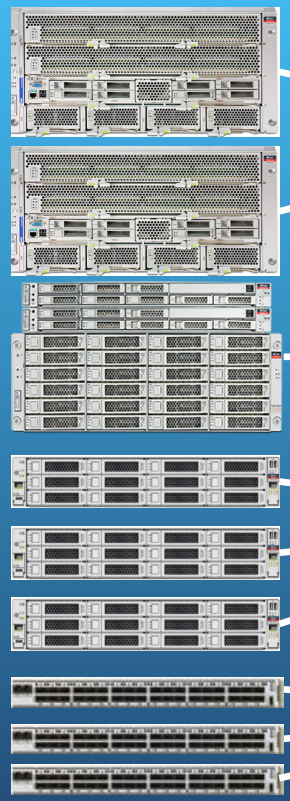
Compute Nodes - RAC

Storage Servers – ASM Redundancy

Infiniband Switches

enkitec

Redundant Hardware – SPARC Supercluster



Compute Nodes - RAC

ZFS Storage Appliance - RAID

Storage Servers – ASM Redundancy

Infiniband Switches

enkitec

Exadata Cell Failures

ORACLE
EXADATA

Critical: Hardware Alert 379_1

Event Time 2012-08-22T00:36:58-05:00

Description Hard disk status changed to predictive failure.

Status	PREDICTIVE FAILURE
Manufacturer	HITACHI
Model Number	H7220AA30SUN2.0T
Size	2.0TB
Serial Number	[redacted]
Firmware	JKAOA28A
Slot Number	10
Cell Disk	CD_10_enkcel01
Grid Disk	RECO_CD_10_enkcel01, DBFS_DG_CD_10_enkcel01, DATA_CD_10_enkcel01

Affected Cell

Name	enkcel01
Server Model	SUN MICROSYSTEMS SUN FIRE X4275 SERVER SATA
Chassis Serial Number	1017XFG056
Release Version	11.2.3.1.1
Release Label	OSS_11.2.3.1.1_LINUX.X64_120607

Recommended Action The data hard disk has entered predictive failure status. It will soon fail and should be replaced at the earliest opportunity. A white cell locator LED has been lit to help locate the affected cell, and an amber service action LED has been lit on the drive to help locate the affected drive.

enkitec

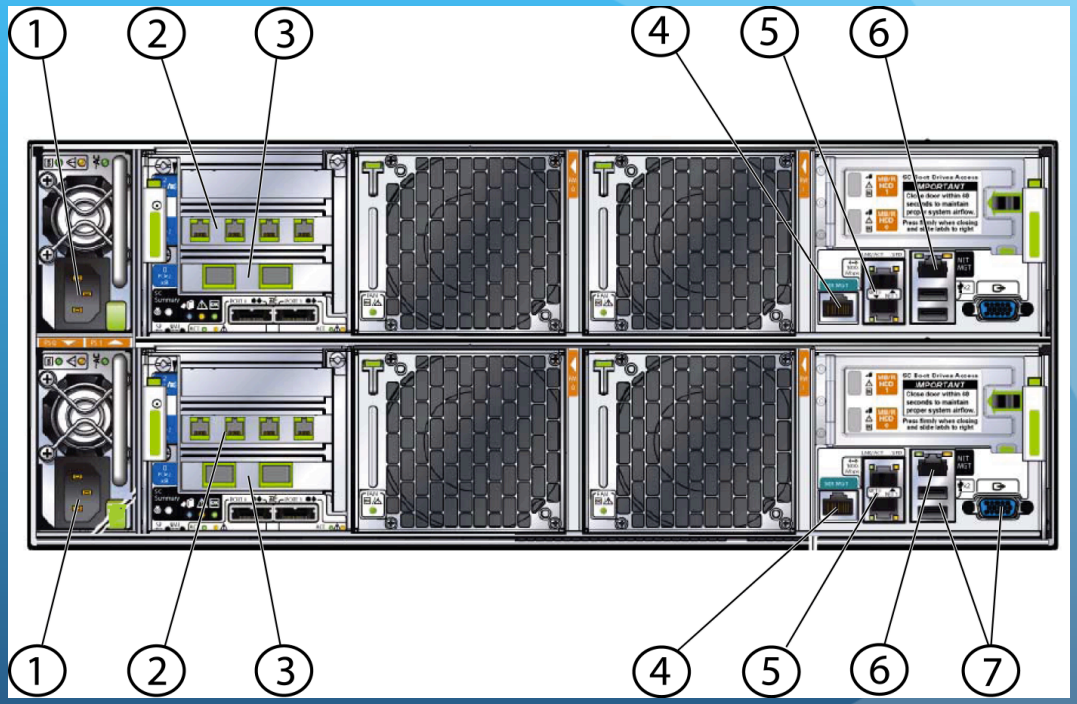
Redundant Hardware – ODA

- Multiple Server Nodes
  - Oracle RAC
  - Multiple Disk Controllers per Server Node
  - Dual-ported SAS disks
- Multiple Operating System Disks
  - Software RAID

enkitec

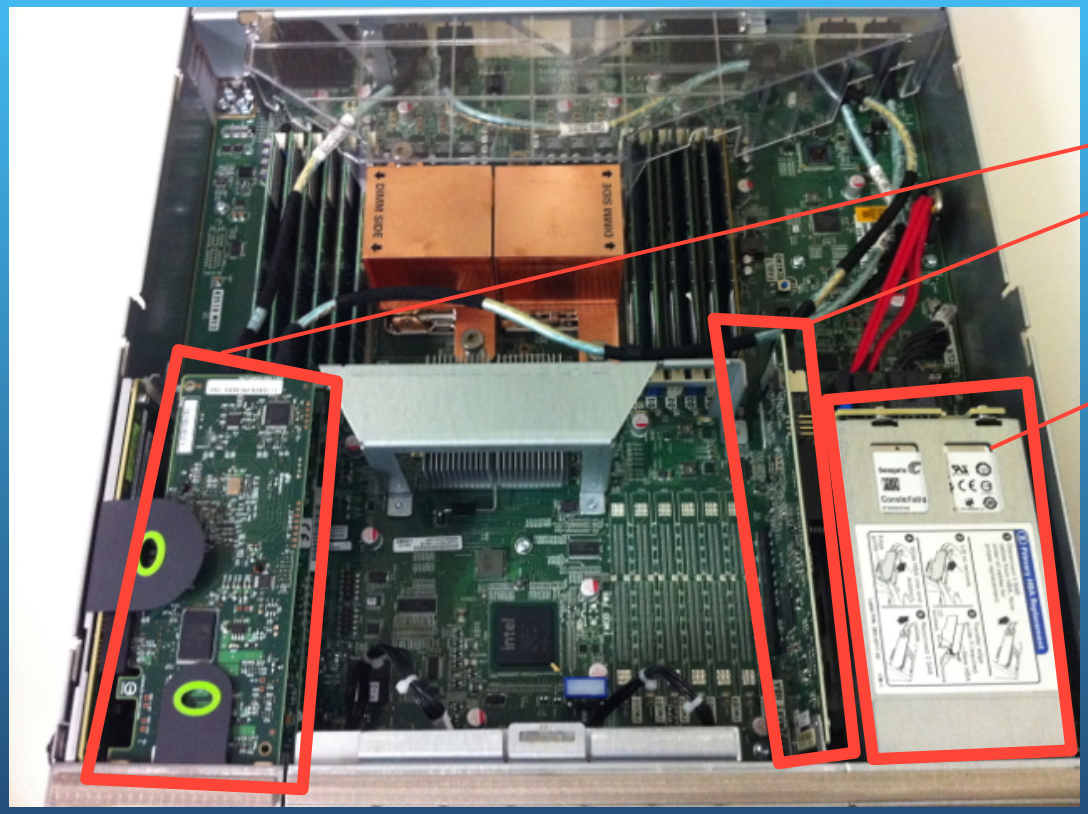
Redundant Hardware – ODA

1. Power
2. (4) 1GbE Ports
3. (2) 10GbE Ports
4. ILOM Serial
5. (2) 1GbE Ports
6. ILOM Network
7. USB/VGA Ports



enkitec

Redundant Hardware – ODA

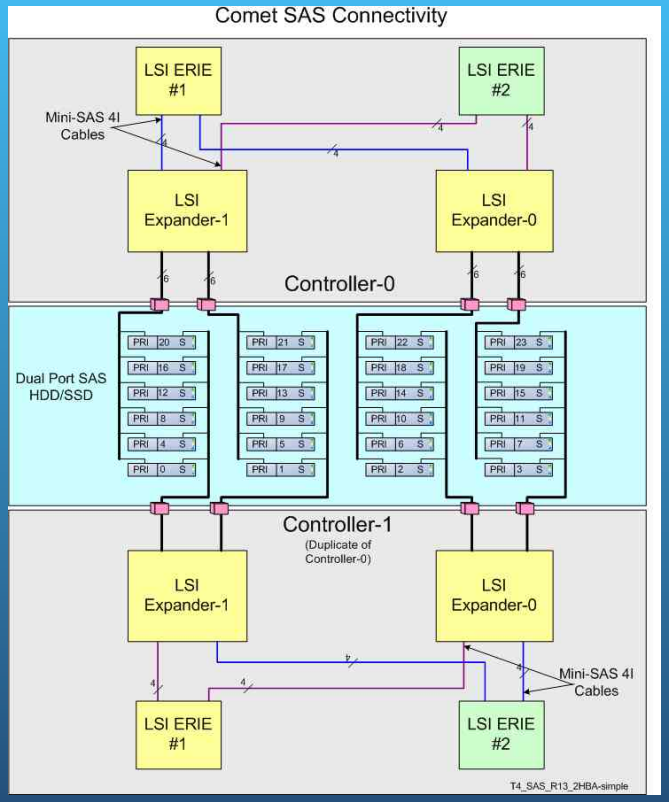


Dual RAID Controllers

Dual Operating System Disks

enkitec

Redundant Hardware – ODA



- Multiple RAID Controllers
- Linux Multipathing

```
multipaths {
    multipath {
      wwid 35000c5003a446893
      alias HDD_E0_S01_977561747
      mode 660
      uid 1000
      gid 1006
}
```

enkitec

ASM Redundancy

- High or normal redundancy
  - High creates 3 copies of each block
  - Normal creates 2 copies of each block
- Cells are arranged into “fail groups”
  - No 2 copies of a block are placed in the same fail group
  - Each cell is a fail group

enkitec

ASM Redundancy - Normal

AV	HW	EK	PR
JU	CT	LN	OS

BX	AE	DJ	FL
IQ	GH	MO	PS

BC	DW	FM	GT
IR	QX	NV	KU

- When an extent is written, the secondary copy is written to one of 8 partner disks

enkitec

ASM Redundancy - High

CI	AL	DN	HP
EO	GJ	FK	BM

AO	FJ	BL	HN
DP	CK	EI	GM

FK	DI	AP	EL
GJ	CM	BN	HO

- When an extent is written, the secondary copy is written to 2 of 8 partner disks
- Still 8 partner disks
- On quarter rack, 1 copy of everything* on each cell

enkitec

ASM Redundancy – Disk Failures

- Each Diskgroup has disk_repair_time Attribute

```
SYS:+ASM1> l
  1  select g.name "Diskgroup", a.name "Attribute", a.value "Value"
  2  from
  3    v$asm_diskgroup g, v$asm_attribute a
  4  where
  5    a.group_number=g.group_number
  6  and
  7    a.name like '%repair%'
  8* order by 1,2
SYS:+ASM1> /

Diskgroup  Attribute                Value
---------  ---------------------    -------
DATA       disk_repair_time         3.6h
DBFS_DG    disk_repair_time         3.6h
RECO       disk_repair_time         3.6h
```

enkitec

ASM Rebalance

✗	HW	EK	PR
JU	CT	LN	OS

BX	AE	DJ	FLV
IQ	GH	MO	PS

BC	DW	FM	GT
IR	QX	NVA	KU

After `disk_repair_time` reaches 0, the disk is dropped and a rebalance is initiated

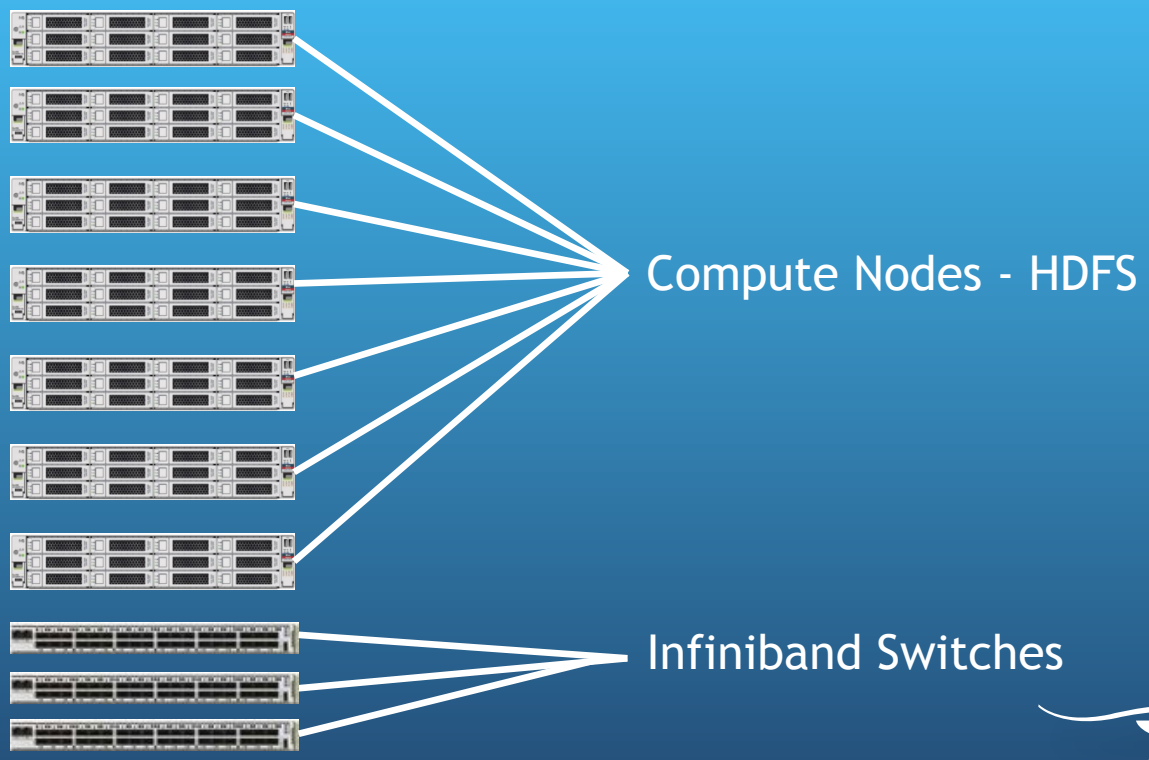
enkitec

And Now For a Demo…

- ACTEST diskgroup simulates full rack
- 168 disks in 14 failgroups
- Mapped partner disks
- We'll "pull" a few drives and see what happens

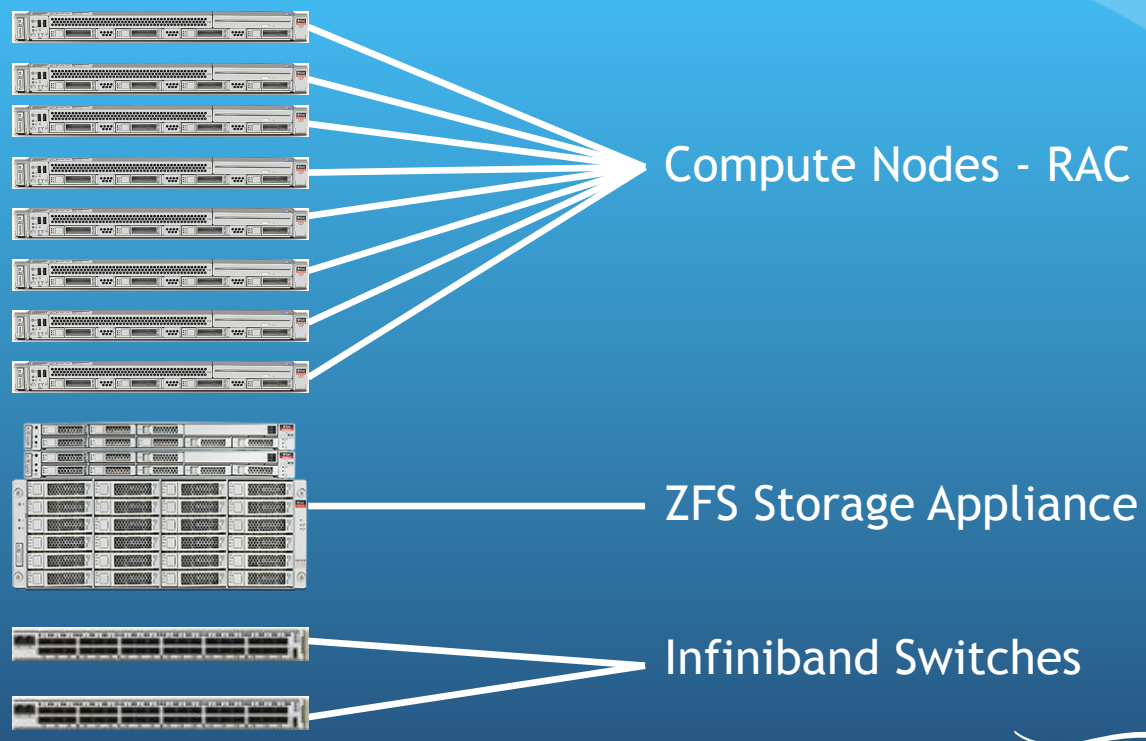
enkitec

Redundant Hardware – Big Data Appliance



enkitec

Redundant Hardware - Exalogic



enkitec

Questions?

Contact Information: Andy Colvin

email - andy.colvin@enkitec.com

web - http://www.enkitec.com

blog – http://blog.oracle-ninja.com

twitter - @acolvin

A Member Of The Inc. 5000

ORACLE Platinum Partner
Specialized
Exadata

enkitec